

UNDERSTANDING THE MYB GENE FAMILY

A Thesis

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Master of Science

by

Kaileigh D. Ahlquist

May, 2018

© 2018 Kaileigh D. Ahlquist

ABSTRACT

MYB transcription factors (TFs) serve many important regulatory roles in plants, making them an important topic of study. However, like many large gene families, there are many inherent difficulties in understanding both the individual roles, and evolutionary history of MYB TFs. Individually, it may be difficult to isolate the functions of a given MYB TF, as there are numerous similar TFs serving related or overlapping functions. As a group, the similarities between members of a gene family can make it difficult to disentangle the forces causing their propagation and maintenance. There are many essential pathways governed by TFs that have evolved via massive expansions, such that a high degree of similarity and redundancy occurs within the gene family. This thesis is concerned with confronting this complex issue, attempting to learn as much as possible about the history and patterns found within the MYB TF family. The objectives of this research are to examine the results of screens involving MYB TF binding partners, and to apply new tools and methods to MYB genes to learn about their evolutionary history. Based on the meta-analysis of biological screens and the use of phylogenetic and recombination detection techniques, this research proposes mechanisms to explain patterns observed and improve the predictive power of computational approaches.

BIOGRAPHICAL SKETCH

Kaileigh Ahlquist was born in Providence, RI in 1989. Ahlquist graduated from Classical High School in 2007, and from Reed College in 2011. After being accepted to Cornell in the Department of Plant Breeding and Genetics in 2011, Ahlquist was awarded the Presidential Life Sciences Fellowship for the 2011-2012 academic year. The Fellowship provided support for research and field work experiences in the labs of Dr. Mark Sorrells, Dr. Rebecca Nelson, Dr. Owen Hoekenga and Dr. Phillip Griffiths. Kaileigh Ahlquist joined the lab of Dr. Walter DeJong in September 2012. While at Cornell, Ahlquist served the Department of Plant Breeding and Genetics as a representative for the Graduate and Professional Student Association from 2012-2014. Ahlquist's research interests continue to center on the intersection of genetics, computational methods, and applications of research for improving human life.

ACKNOWLEDGMENTS

I would like to thank Dr. Walter De Jong and Dr. Jeffrey Doyle for their advice and editing over the course of producing this manuscript. They have stuck with me through many deadlines and obstacles, and provided excellent guidance throughout my degree. I would also like to thank my committee members Dr. Stephen Reiners and Dr. Robin Bellinder, and Director of Graduate Studies Dr. Michael Mazourek, all of whom provided vital assistance in helping me reach this point. Many more people associated with the Section of Plant Breeding and Genetics helped me to achieve this degree at various times. I am grateful to all of them, especially my friend Anna Levina who has been part of my experience at Cornell from start to finish.

Over the course of producing this thesis I was lucky enough to marry Joshua Keller.

He has been there all along, providing invaluable support for my ambitions and encouraging me in moments of doubt. I also thank my family, Kathy, Steve, Ayla and Alex Ahlquist, who fill me with pride and inspire me every day.

TABLE OF CONTENTS

ABSTRACT	III
TABLE OF CONTENTS	VI
CHAPTER 1	1
Introduction	
CHAPTER 2	3
Structure of MYB Genes	
CHAPTER 3	10
Evolution and Classification of MYB Genes	
CHAPTER 4	15
Review of Demonstrated Interactions	
CHAPTER 5	19
Results	
CHAPTER 6	33
Discussion	
CHAPTER 7	37
Materials and Methods	
REFERENCES	43

CHAPTER 1

INTRODUCTION

MYB transcription factors (TFs) have been extensively studied due to the important regulatory roles they serve in plants. MYB TFs control phenylpropanoid biosynthesis, trichome and root hair development, and are involved in cell fate and hormone signaling pathways (Dubos et al. 2010; Zhao et al. 2013; Du et al. 2012). Because MYB genes are numerous in plants, and biological analyses are costly, computational methods are frequently used to make predictions about MYB TF functions. *In silico* classification of MYB genes is commonly used to cut down on the number of experiments necessary *in planta*. For example, if *in silico* classification can be used to predict MYB genes that are likely to have functional overlap, multi-gene knockout combinations can be recommended to create a desired phenotype.

This problem is not limited just to the MYB family of TFs, but is a common issue for large gene families. Within large families there are many examples of functional redundancy, or instances where genes cannot be targeted specifically due to similarities with other members of the gene family. Gene phylogenies are frequently used to inform the design of experiments, with the intention of identifying highly similar genes that have functional overlap or cause off-target interactions. Yet, the ability of these methods to accurately predict functional overlap or off-target

interactions has not been tested. The objectives of this research were to examine the results of screens involving MYB TF binding partners, applying new tools and methods to MYB genes to learn about their evolutionary history. Based on the meta-analysis of biological screens and the use of phylogenetic and recombination detection techniques, this research proposes mechanisms to explain patterns observed and improve the predictive power of computational approaches. The results suggest that researchers should consider specific functional regions when searching for genes that may have related activity. By developing new computational methods researchers can improve candidates for *in vitro* and *in vivo* experiments.

CHAPTER 2

STRUCTURE OF MYB GENES

MYB genes, named for the discovery of the v-MYB transcriptional regulator in avian myeloblastosis virus (Klempnauer et al. 1982), encode transcription factors containing one or more characteristic MYB motifs. The MYB motif is typically 50-53 amino acids long and is characterized by three helices and a helix-turn-helix conformation. This structure has been verified by NMR and X-ray crystallography (Ogata et al. 1994). MYB motifs facilitate transcription by binding with DNA, and sometimes by binding with other proteins as well. Many MYB TFs contain multiple copies of the MYB motif. When this occurs, the MYB TF is often labelled by the number of MYB motifs. For example, a MYB TF with one MYB motif may be called a 1R- (or “one repeat”) MYB, while a MYB TF with five MYB motifs may be called a 5R-MYB. In addition, the relationship of individual MYB motifs to motifs in 3R-MYBs is often noted. The repeated motifs in 3R-MYBs are labelled repeat 1 (R1), repeat 2 (R2) and repeat 3 (R3). Most 2R-MYBs contain motifs similar in sequence to R2 and R3, and so are frequently designated as R2R3-MYBs, to differentiate them from other 2R-MYBs whose motifs may have other origins.

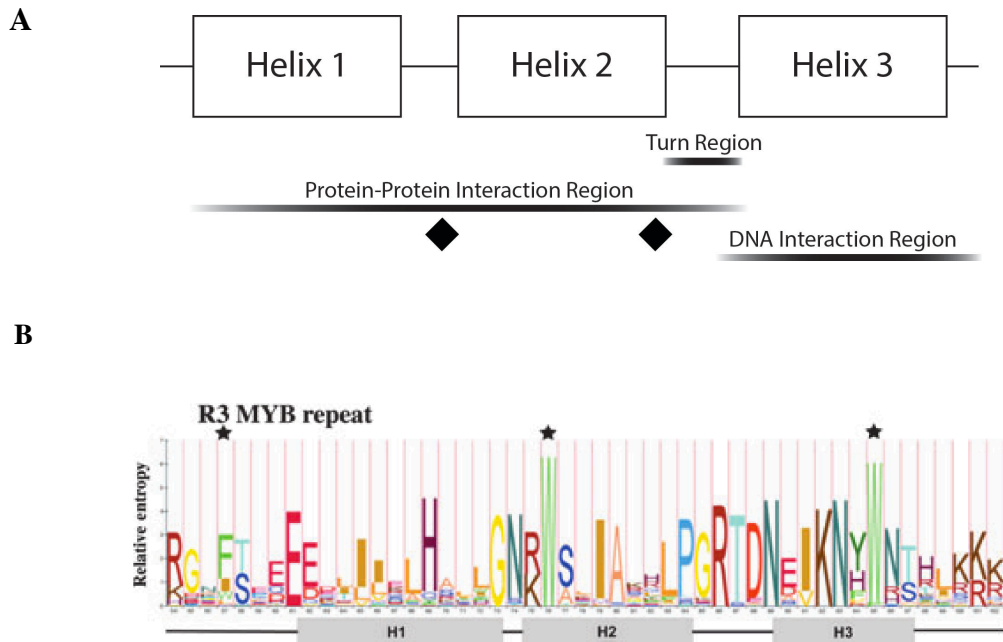


Figure 1. Features of MYB motifs.

- A. Structure of a MYB motif.** All MYB motifs contain three helices and a turn region while the DNA and Protein-Protein Interaction Regions shown are present only in some instances. The DNA Interaction Region associated with the third helix has been found only in R2 and R3 versions of the MYB motif. DNA binding domains have also been observed outside the third helix in less common instances of the MYB motif (Hwang et al. 2001). To date the Protein-Protein interaction region has been found only in the R3 MYB motif. Diamonds mark amino acid residues 20 and 33 of the MYB motif, residues that had the greatest impact on protein-protein interactions in a study by (Zimmermann et al. 2004).
- B. Sequence logo of R3 MYB motif** (from Feller 2011). Stars indicate conserved tryptophan residues characteristic of the MYB motif. The logo was generated with a Pfam seed alignment containing 155 MYB genes, using the method described by Schuster-Boeckler et al. 2004, available at <http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi>. H1, H2 and H3 at bottom indicate location of helix 1, helix 2 and helix 3, respectively.

Protein-Protein Interactions

MYB motifs can contain domains responsible for interactions with other MYB proteins as well as with other protein classes like bHLH and WD40 (Zimmermann et al. 2004; Zhao et al. 2013; Montefiori et al. 2015). One such site, verified by truncation and mutation, starts at the first helix of the R3 MYB motif and ends at the turn region located between the R2 and R3 helices. The two residues with the greatest influence on MYB-bHLH interaction are located on opposite edges of Helix 2 (Figure 1A). This sub-motif is not found in all MYB genes known to interact with bHLH partners, although it is possible that different sequences in the same region could be responsible for other MYB-bHLH interactions.

Most MYB and bHLH TFs that have been extensively studied appear to interact strongly with just a few partners, and have little or no interaction with other tested partners (Zimmermann et al. 2004; Zhao et al. 2013; Lumba et al. 2014, Montefiori et al. 2015). The specificity of binding between partners is not necessarily exclusive, as one MYB TF may bind with three or four bHLH TFs, and vice versa. Pull-down experiments tend to identify a handful of distinct partners, not the dozens we would expect if there was fully promiscuous binding between members of the two very large TF families (Zimmermann et al. 2004; Lumba et al. 2014).

DNA Binding

The third helix of a MYB motif is sometimes referred to as the DNA-recognition helix. Changes to this region typically disrupt or alter DNA binding. When two or more MYB motifs are part of a single MYB TF, DNA binding usually involves (and often requires) more than one MYB motif. In these cases, each motif recognizes and binds a portion of the TF-binding site, and the adjacent MYB motifs appear to stabilize binding (Ogata et al. 1994; Ogata et al. 1996; Jia, Clegg, and Jiang 2004).

MYB TFs have strong affinities for certain sequences or classes of sequences. While NMR and other visualization and modeling techniques have demonstrated how two MYB motifs work together to bind DNA, less is known about how MYB TFs with a single MYB motif bind DNA. Some studies of 1R-MYB TFs have identified DNA-binding residues in the first and second helices, which differs from the protein-protein interaction role these helices serve in MYB TFs with multiple motifs (Hwang et al. 2001; Prouse and Campbell 2012).

A comparison of all published binding sites for MYB TFs has revealed a great deal of sequence diversity. Some plant MYB TF binding sites are similar to those found in mammals and other eukaryotes, but there are many MYB TF binding sites unique to plants (Prouse and Campbell 2012). The diversity of plant MYB genes has given rise to a number of interesting phenomena not found with MYB genes in other organisms. These include MYB-DNA interactions that are modulated by chemical conditions that alter MYB binding (Koshino-Kimura et al. 2005; Serpa et al. 2007), or interactions

that involve the competition of MYB TFs for the same binding sites (Feldbrügge et al. 1997; Liao et al. 2008; Lu et al. 2002).

MYB Motifs

Because many MYB genes contain two or more copies of the MYB motif, a given motif is typically labeled as to which repeat it most resembles. MYB genes have a variety of structures, with differing placement and even ordering of MYB motifs (Figure 2).

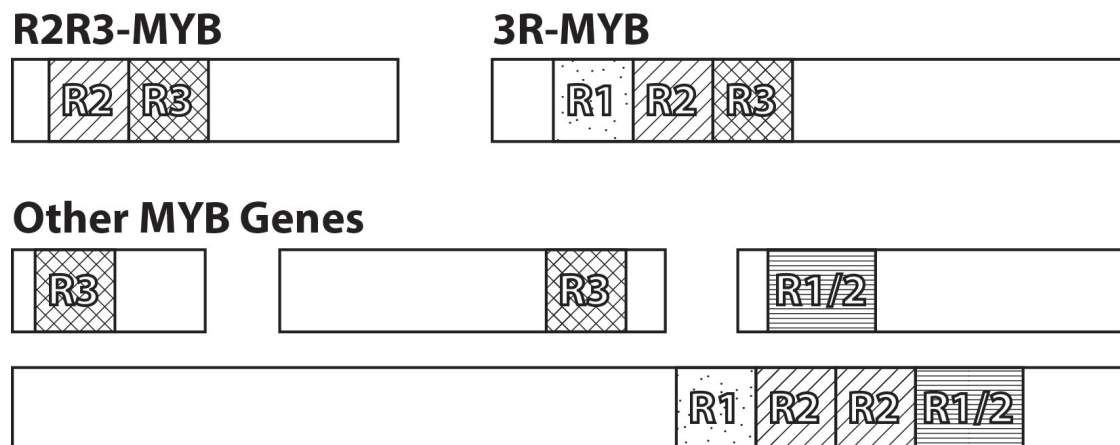


Figure 2. MYB gene structures.

All R2R3- and 3R- MYB genes match the general structures shown. Other MYBs and MYB-related genes are extremely diverse, with the structures shown representing just a fraction of the diversity of this group. These example structures are drawn from Dubos et al. 2010, Du et al. 2013 and direct observation.

R2R3-MYB Genes

While R2R3-MYB genes are found in all known eukaryotes, the subfamily has expanded dramatically in plants, with most plants having over 100 R2R3-MYB genes (Feller et al. 2011).

3R-MYB Genes

3R-MYB genes contain R2 and R3 motifs similar to those found in R2R3-MYBs, preceded by the R1 motif. 3R-MYB genes are also found in all eukaryotes, but unlike the R2R3-MYB genes, their numbers have not drastically increased in plants relative to other eukaryotic organisms. Most eukaryotes, including plants, have less than ten 3R-MYB genes (Feller et al. 2011).

Other MYB Genes

In addition to genes readily classified as either 3R-MYBs or R2R3-MYBs, there are additional MYB genes that carry between 1 and 5 repeats of the MYB motif. These motifs can often be classified as being most closely related to R1, R2 or R3. Some motifs represent special cases, such as the R1/2 motif, sometimes cited as the progenitor of the R1 and R2 motifs (Romero et al. 1998; Feldbrügge et al. 1997). While the R1/2 motif does bear similarities to R1 and R2, descent is not the only possible explanation, and would be difficult to verify.

The structures of other MYB genes suggests that they may have evolved by the insertion, deletion and/or mutation of MYB motifs from the common R2R3 form

(Feller et al. 2011). There is a great deal of structural diversity for MYB genes outside the R2R3- or 3R- categories, with the location of the MYB motif in the gene being especially variable (Du et al. 2013). While some of these genes may be pseudogenes, testing for expression of 127 single-repeat MYB genes in soybean found that just 22 genes were not expressed, while others were expressed across many tissues (Du et al. 2013). The same authors demonstrated similar results in maize. MYB genes with a single repeat have increased in number in angiosperms relative to other eukaryotes, with typical angiosperms having between 40 and 100 copies of these genes (Du et al. 2013).

CHAPTER 3

EVOLUTION AND CLASSIFICATION OF MYB GENES

MYB Gene Proliferation

MYB TFs are found across all eukaryotic lineages, but the family has dramatically expanded in plants. Expansion of the MYB TF family has been demonstrated in the moss *Physcomitrella patens* (Shapiro et al. 2008), with significant further expansion occurring within the angiosperm lineage, prior to the divergence of monocots and dicots (Rabinowicz et al. 1999; Du et al. 2013; Bedon et al. 2010).

The mechanisms of MYB gene duplication are partially understood. Whole genome duplication, chromosome duplication and major rearrangements, followed by preferential retention of duplicated MYB genes and other transcription factors, have all played a role (Shiu et al. 2005; Blanc et al. 2000; Riechmann et al. 2000; Seoighe and Gehring 2004; Cannon et al. 2004). Tandem duplications have likely also played a role (Matus et al. 2008; Cannon et al. 2004; Yanhui et al. 2006). Transposition of MYB elements via recombination events has also been proposed as a possible mechanism either to increase the number of MYB genes, or to maintain similarity of MYB genes through partial or total gene conversion (Feller et al. 2011; Woodhouse, et al. 2010).

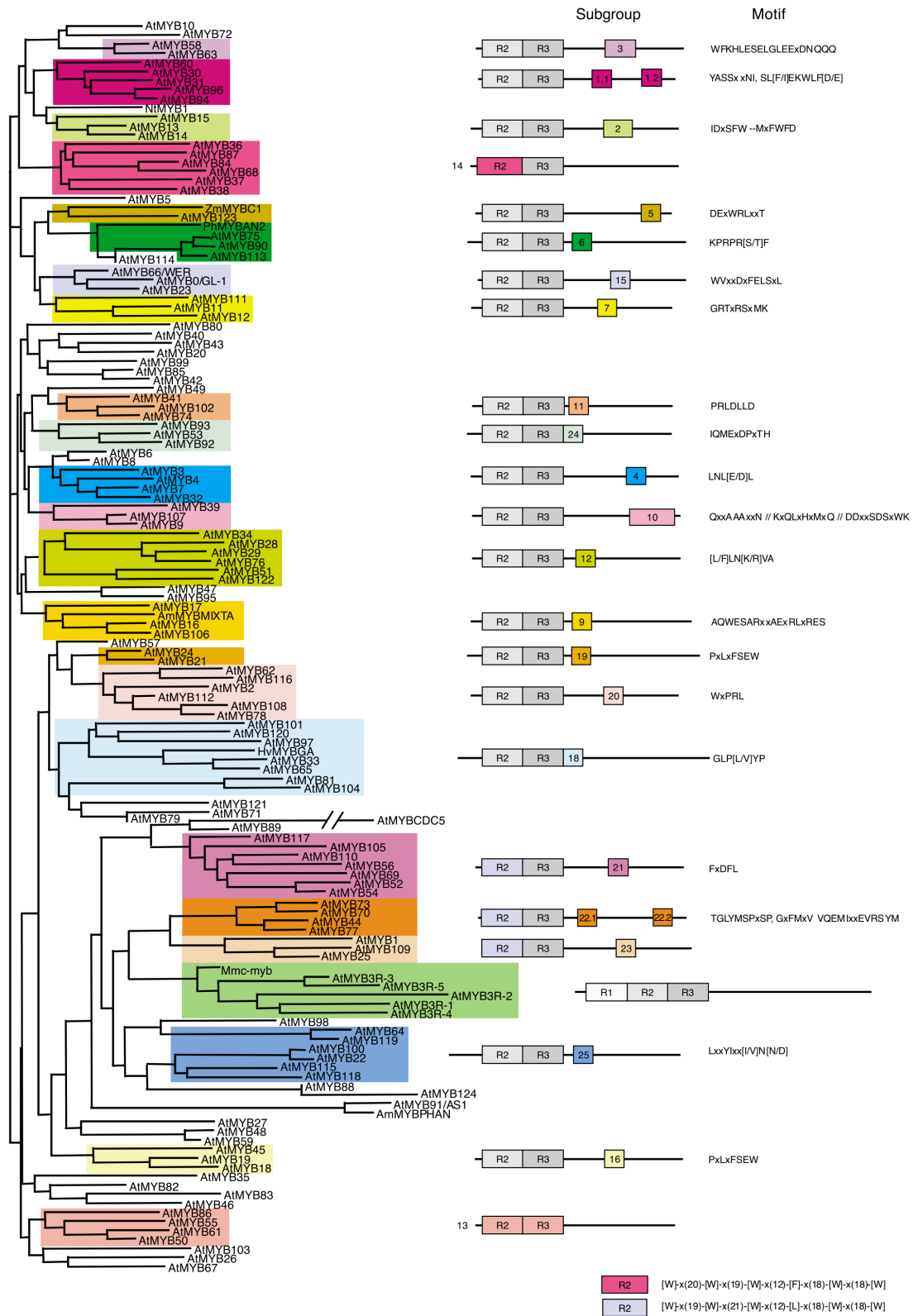


Figure 3. Classification of Arabidopsis MYB TFs from Stracke et al. 2001

Figure from Stracke et al. 2001 showing alignment of amino acid sequences: “Relationship of *A. thaliana* MYB proteins that have two or three repeats. The AtMYB factors were clustered using PHYLIP, and motifs were detected using MEME. Subgroups were designated as previously reported (Kranz et al. 1998) however, some amino-acid motifs were newly interpreted, additional entries were integrated using MEME, and two new subgroups were added because of the increased data set. Some subgroups defined before are not apparent because predominantly MYB genes from *A. thaliana* were considered.” Classification based on this scheme was updated with additional MYB TFs in Dubos et al. 2010.

Current Techniques

Some of the most prominent work classifying MYB genes was performed by Kranz et al. 1998 and expanded upon by Stracke et al. 2001 and Dubos et al. 2010 (Figure 3). Their classification schemes use phylogeny to group together related R2R3 and 3R MYB genes, leading to prediction of function based on phylogenetic position and similarity to genes with known function. The same techniques have been used by others to extend research performed in the model organism *Arabidopsis* into less-researched crops including sugar beet, Chinese Cabbage, soybean, cucumber and more (Du et al. 2012a; Du et al. 2012b; Hou et al. 2014; Zhou et al. 2015; Cao et al. 2013; Katiyar et al. 2012; Stracke et al. 2014). However, little research has addressed whether the phylogenetic relationships can be used to accurately predict what DNA or other proteins MYB TFs will bind.

Closely related MYB TFs do not necessarily interact with the same DNA and protein partners, and experiments that include broad testing of binding partners has captured

MYB and MYB-like genes that appear to be distantly related based on a phylogenetic tree (Lumba et al. 2014; Taylor-Teeples et al. 2015). Research where the presumed phylogenetic relationships were used as criteria for testing does show closely related genes sharing binding partners (Frerigmann et al. 2014; Zimmermann et al. 2004). However, broader biological testing tends to reveal a more complex network of binding partners that suggests the scope of current *in silico* methods may be too narrow (Zimmermann et al. 2004).

Predictive Ability of MYB TF Phylogenies

Existing datasets make it possible to compare the relationships observed in a phylogeny with demonstrated molecular interactions. BIOGRID is a protein interaction database that includes 30 *Arabidopsis* MYB TFs found to interact with 31 bHLH TFs across 20 publications (Chatr-aryamontri et al. 2015). Similarly, for protein-DNA interactions, the *Arabidopsis* Gene Regulatory Information Server (AGRIS) maintains the data file AtRegNet, which records known and predicted interactions between *Arabidopsis* TFs and DNA sequences (Yilmaz et al. 2011). At the time of writing, this database reported 27 unique MYB TFs that have been confirmed to pair with 93 unique gene targets. Discovery of these interactions spanned 13 publications. For both these datasets, the interactions of most interest for the current study are those discovered through untargeted screens, i.e. screens against large, comprehensive pools of unselected interactors.

While all known MYB interactions were used in initial analyses, publications that used untargeted screens and reported four or more MYB TFs interacting with the same

partners were selected for additional analyses. Papers that fell into this category were Lumba et al. 2014 and Taylor-Teeple et al. 2015. Also considered was Zimmermann et al. 2004, which used a combination of untargeted screens and prior phylogenetic information. Methods used in these papers are described in detail below.

CHAPTER 4

REVIEW OF DEMONSTRATED INTERACTIONS

Zimmermann (Zimmermann et al. 2004)

Interacting proteins in this paper were identified using a mixture of biased and untargeted methods. MYB TFs to test were initially selected on the basis of their membership in groups 5, 6 and 15 as designated by Stracke et al. 2001. The yeast two-hybrid system was used to test protein-protein interactions between the selected genes and the bHLH genes EGL3 (AT1G63650), bHLH012 (AT4G00480) and TT8 (AT4G09820). MYB proteins were fused to the GAL4 activation domain and bHLH proteins were fused to the GAL4 DNA binding domain. The reciprocal experiment could not be performed due to the inherent activating ability of the MYB TFs. The selection of these genes using the Stracke group designation is biased against the discovery of distantly related genes.

Untargeted results were obtained by using a pull-down assay with bHLH genes as bait against activation domain-fused cDNA libraries. The pull-down assay confirmed the results of the yeast two-hybrid assay by identifying all of the MYB genes shown to interact with a given bHLH, and also identified additional genes that interacted with each bHLH. The additional genes found through untargeted methods are highlighted with a star in Table 1 (Zimmermann et al. 2004). AT1G71030.1 is the only additional gene that contains an identifiable MYB motif according to PROSITE. The other

starred genes contain conserved sequences that align with MYB motifs, but the similarity is not sufficient to be detected by PROSITE.

Table 1. Genes encoding MYB TFs found to interact with bHLH TFs AT1G63650, AT4G00480, AT4G09820 (protein-protein)

MYB TF	Alias	AT1G63650 (EGL3)	AT4G00480 (bHLH12)	AT4G09820 (TT8)	Stracke Classification
AT3G27920.1	GL1	X	X		15
AT1G66380.1	AtMYB114	X	X	X	6
AT5G40330.1	TT2	X	X	X	15
AT5G14750.1	WER	X	X		15
AT1G66390.1	PAP2	X	X	X	6
AT3G13540.1	AtMYB5	X	X	X	NA
AT1G56650.1	PAP1	X	X	X	6
AT1G66370.1	AtMYB113	X	X	X	6
AT5G35550.1	TT2	X	X	X	5
AT1G71030.1*	AtMYBL2	X	X	X	NA
AT2G46410.1*	CPC	X	X	X	NA
AT1G01380.1*			X		NA
AT2G30420.1*			X		NA

Lumba (Lumba et al. 2014)

A total of 282 genes whose expression changed at least 2-fold relative to wildtype in an ABA deficient mutant were tested for interaction with bHLH TF AT1G10585 using a yeast two hybrid assay (Lumba et al. 2014). Four MYB TFs were found to interact with bHLH TF AT1G10585.

Table 2. MYB TFs interacting with bHLH TF AT1G10585 (protein-protein)

MYB TF	Alias	Stracke Classification
AT5G54230.1	MYB49	NA
AT3G50060.1	MYB77	22
AT4G05100.1	MYB74	11
AT2G47460.1	MYB12	7

Taylor-Teebles (Taylor-Teebles et al. 2015)

Promoter sequences for 50 genes known to be important for xylem cell specification were screened for interaction with 467 transcription factors expressed in root xylem using the yeast one hybrid system. One of the promoters tested was for AT2G30490, a gene that encodes cinnamate-4-hydroxylase (Taylor-Teebles et al. 2015). Nine different MYB transcription factors interacted with the AT2G30490 promoter to drive expression of a reporter gene (Table 3).

Table 3. MYB TFs interacting with the promoter of AT2G30490

MYB TF	Alias	Stracke Classification
AT5G17800.1	AtMYB56	21
AT3G08500.1	MYB83	NA
AT4G22680.1	MYB85	NA
AT1G16490.1	MYB58	3
AT5G26660.1	ATMYB4	13
AT1G22640.1	MYB3	4
AT1G66230.1	MYB20	NA
AT1G79180.1	MYB63	3
AT2G47460.1	PFG1	7

CHAPTER 5

RESULTS

Classification of MYB Genes

The phylogenetic tree of 215 documented *Arabidopsis* MYB genes in Figure 4 shows that the commonly used MYB TF classification groups of Stracke et al (Figure 3) are maintained as monophyletic groups. However, when annotations of verified interactions are overlaid on this phylogeny (Figure 4), it becomes clear that MYBs that bind to any single interactor can be widely dispersed across the phylogeny. Some interactions from Zimmermann et al. 2004 are tightly clustered, but the interactions tested were not random. Four genes found in the untargeted screen conducted by Zimmerman et al. 2004 would not be detected using an analysis of a MYB gene phylogeny. They contain a segment of the MYB motif responsible for protein-protein interactions, but the motif is not complete enough to detect using PROSITE or other computational means of motif detection. bHLH interactions detected by Lumba et al. 2014 and cinnamate-4-hydroxylase promoter interactions detected by Taylor-Teeple et al. 2015 are spread across the phylogeny with no clearly discernible pattern. This suggests that current phylogenetic screening and classification methods do not tend to predict protein-protein or protein-DNA binding interactions between MYB TFs and their partners.

● Lumba
 ▲ Taylor-Teeple
 ■ Zimmermann

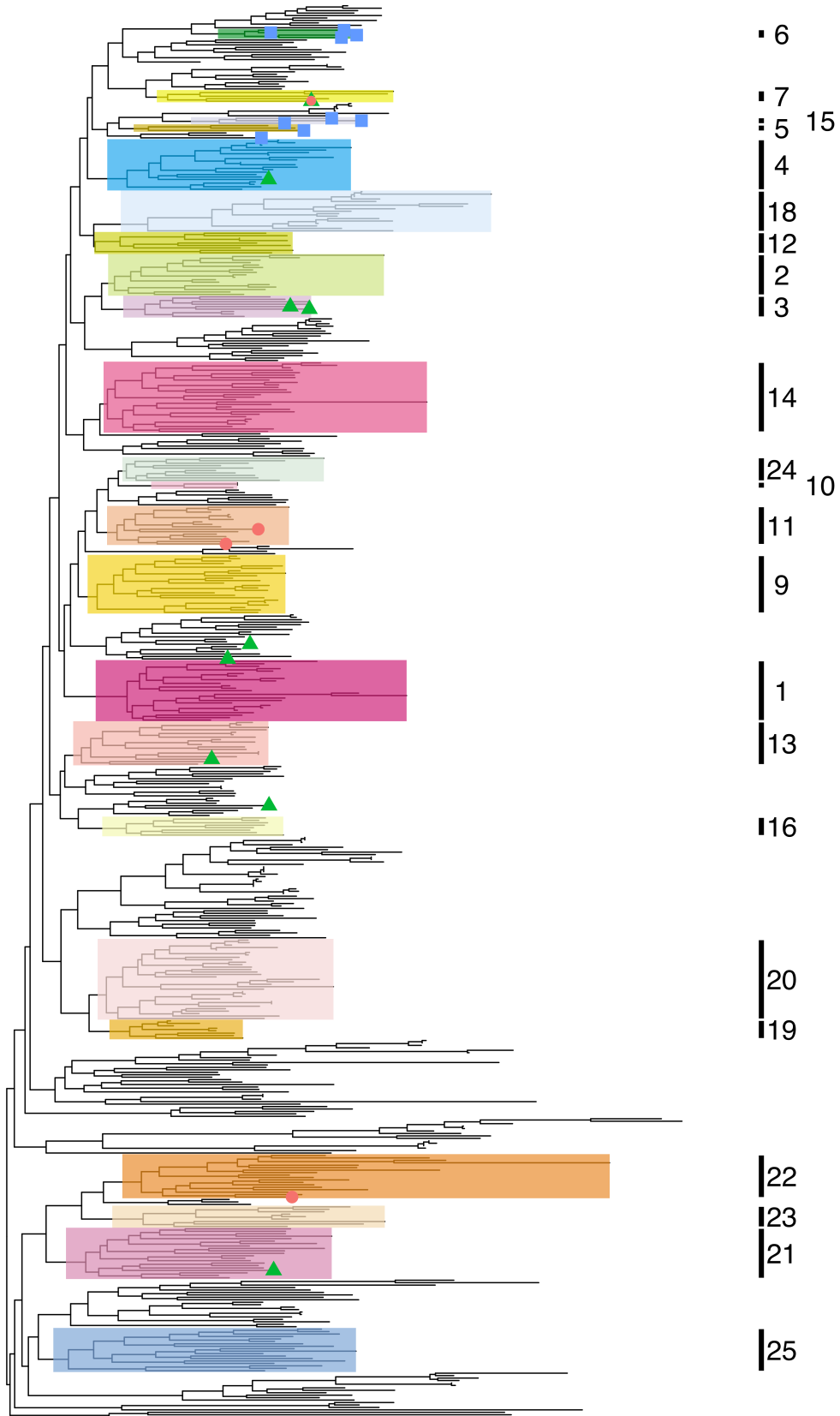


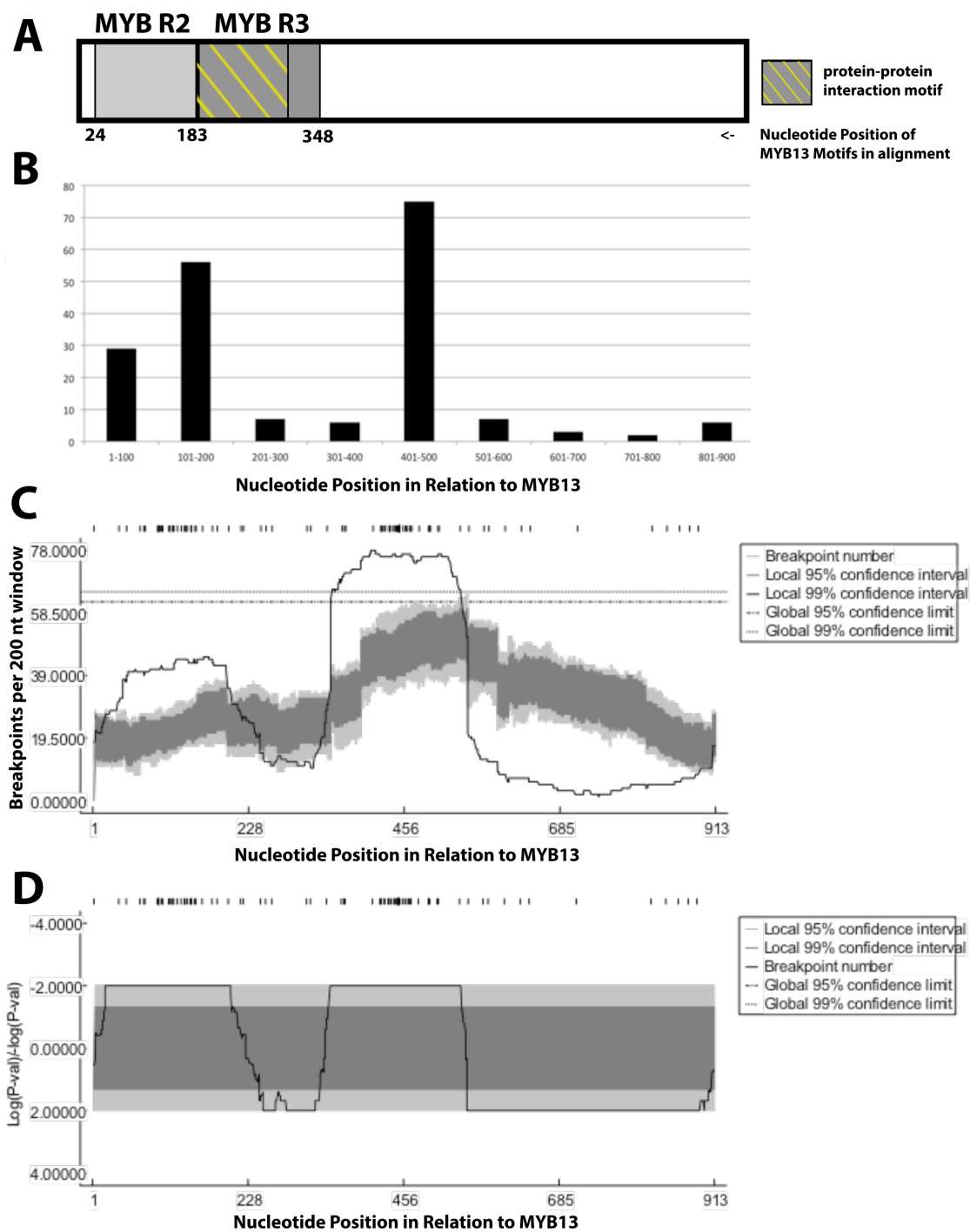
Figure 4. Maximum Likelihood phylogeny of MYB sequences showing Stracke annotations and experimental data.

A Maximum Likelihood phylogeny of MYB protein sequences was generated using RAxML. MYB classifications defined by Stracke et al. 2001 are overlaid by groups of MYB TFs found to interact with the same protein or DNA binding partners. Each MYB group defined by Stracke et al. 2001 is labeled with the same color as in Figure 3, and also labelled with a number to the right. MYB TFs found to interact with a shared protein partner in Zimmermann et al. 2004 are marked with blue squares. These TFs span three of the groups designated by the Stracke et al. classification scheme (groups 5, 6 and 15). MYB TFs found to interact with a shared protein partner in Lumba et al. 2014 are marked with red circles. These TFs span three of the groups designated by the Stracke et al. classification scheme (groups 7, 11 and 22), with an additional MYB TF not classified previously. MYB TFs found to interact with shared DNA binding partners in Taylor-Teeple et al. 2015 are marked with green triangles. These TFs span five of the groups designated by the Stracke et al. classification scheme (groups 3, 4, 7, 13 and 21). A single TF marked with both a green triangle and a red circle was found as an interacting partner in both of the latter two studies.

Recombination in the MYB Gene Family

Both the pattern of the interactions observed and the structural diversity of MYB TFs suggest that recombination may have played a role in the evolution of this gene family. The software package RDP4 (Martin et al. 2015) was used to detect recombination in an alignment of 215 MYB gene nucleotide sequences from *Arabidopsis*. While RDP can be used to look at recombination events within individual genes, the analysis shown in Figure 5 takes a broad view, demonstrating

overall patterns detected within the MYB gene family. Breakpoints indicate the predicted start or end of a recombinant region (shown in Figure 5 B-D). The recombinant segments are shown in Figure 5 E. The locations of the breakpoints are based on the alignment (Ara_MYB_Nuc_1) (see Methods in Chapter 7). MYB13 was arbitrarily used as the reference sequence for the figures below to provide nucleotide locations along the gene sequence. If other genes in the alignment were used as a reference we would expect shifts in the nucleotide locations as a result of indels, but the locations relative to well-conserved gene regions would stay the same. While random recombination events would result in an even distribution of recombinant regions and breakpoints, the data shows recombination events are concentrated in certain regions. Specifically, more putative recombination events were detected in the part of the MYB gene aligning with the R3 MYB motif, and within the conserved motif region in general.



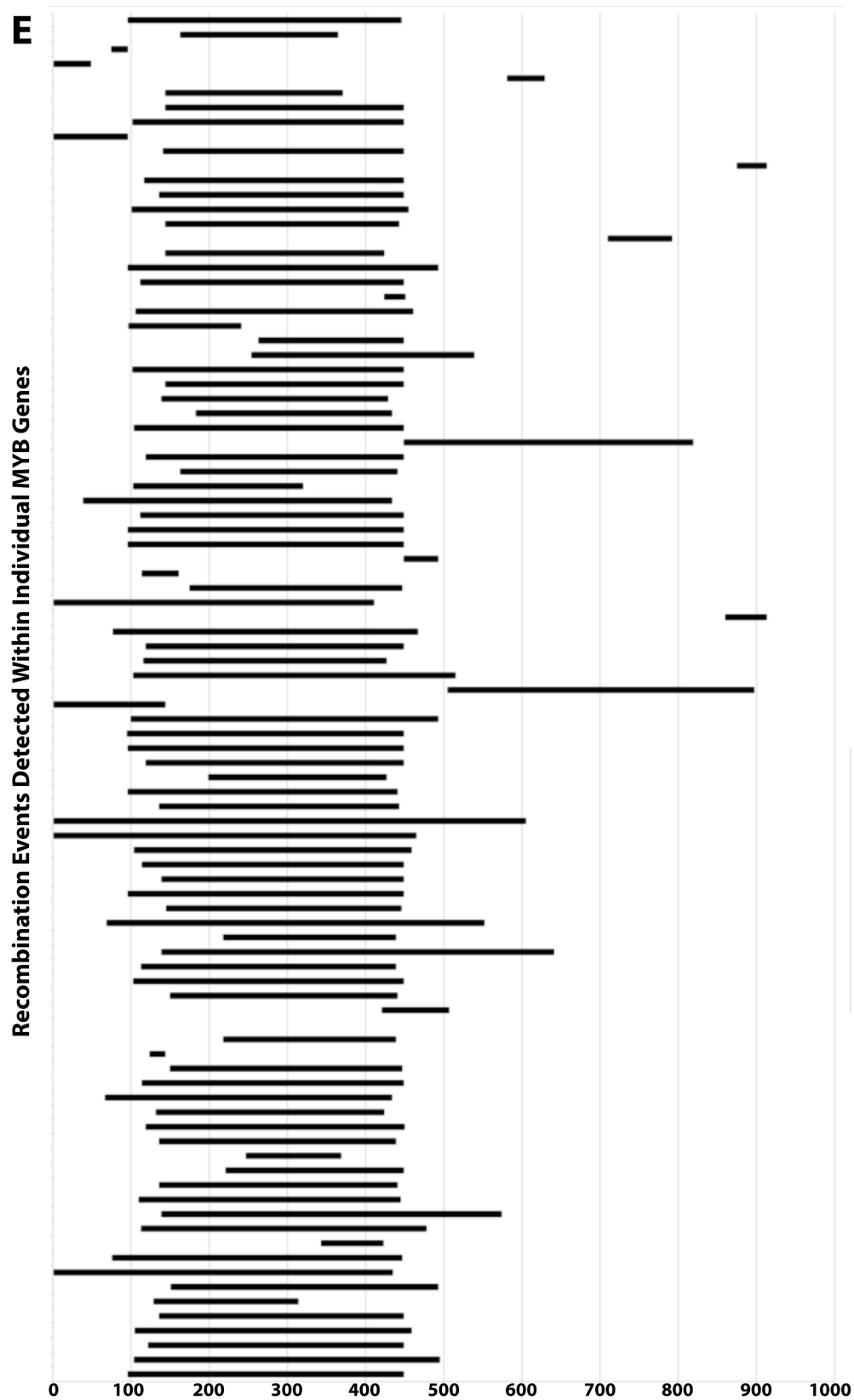


Figure 5. Distribution of detected recombination events across *Arabidopsis* MYB genes. Positions are given as nucleotide position relative to MYB13 in the alignment.

- A.** Diagram of MYB13 with positions of motifs shown. Approximately to scale.
- B.** Histogram of breakpoints detected by RDP.
- C.** The y-axis shows the number of breakpoints detected by RDP4 within a moving 200-nt window, indicated with an unbroken black line. Local confidence estimates (95% in dark gray, 99% in light gray) generated by RDP4 show the relative probabilities of breakpoints across regions of the alignment.
- D.** The y-axis indicates the p-value for the breakpoints detected by RDP4 within a moving 200-nt window. This plot is a transformation of plot C where the confidence intervals (95% in dark gray, 99% in light gray) are held constant and the probabilities of breakpoints, rather than the absolute number of breakpoints, are represented by the unbroken black line.
- E.** Each horizontal segment represents a single MYB gene, with the black bar representing the region of that MYB gene that is predicted to have recombined from another MYB gene in the alignment. This provides a visual summary of all recombination events that are detected between genes in this alignment.

MYB Gene Evolution

Recombination across the MYB R3 repeat and part of the MYB R2 repeat suggest functional significance for these recombination events. For the reference sequence used in Figure 5, the R2 repeat spans positions 24-183 and the R3 repeat spans positions 183-348. The protein-protein interaction motif identified in Zimmermann et al. 2004 encompasses most of the R3 repeat; recombination was detected very frequently at the beginning of this interaction motif, with little recombination occurring at its end. To explore further the evolution of MYB TFs an alignment of protein sequences was created (See Table 6 and Methods in Chapter 7) and divided into subsections representing different TF regions. These sub-alignments were then used to generate the phylogenies shown in figures 6 and 7.

Examining the resulting phylogenies allows us to ask whether all parts of a MYB gene have evolved according to the same pattern. For example, if the clades produced using the first section of the alignment are the same as those produced using the second section of the alignment, this suggests that all parts of the gene have the same evolutionary history. If, as was observed, the phylogenies produced are very different it could suggest distinct evolutionary histories for different sections of the genes. Examples of this can be seen by comparing the phylogenies in Figure 6. Clusters evident in the phylogeny of the whole sequence or the N-terminal sequence are rearranged in the phylogeny drawn from the C-terminal sequence. Focusing on the prominent cluster of genes identified in Zimmerman, we can see in Figure 7 that there are substantial differences in the relationships found when different portions of the

alignment are considered. Drawing a firm conclusion is complicated by the structural diversity of the MYB family, which causes the quality of the alignment to decrease outside of the conserved MYB region. However, we can see in Figure 7 that there is still reasonably strong bootstrap support for the phylogeny built using sequences outside the MYB region.

● Lumba ▲ Taylor-Teeples ■ Zimmermann

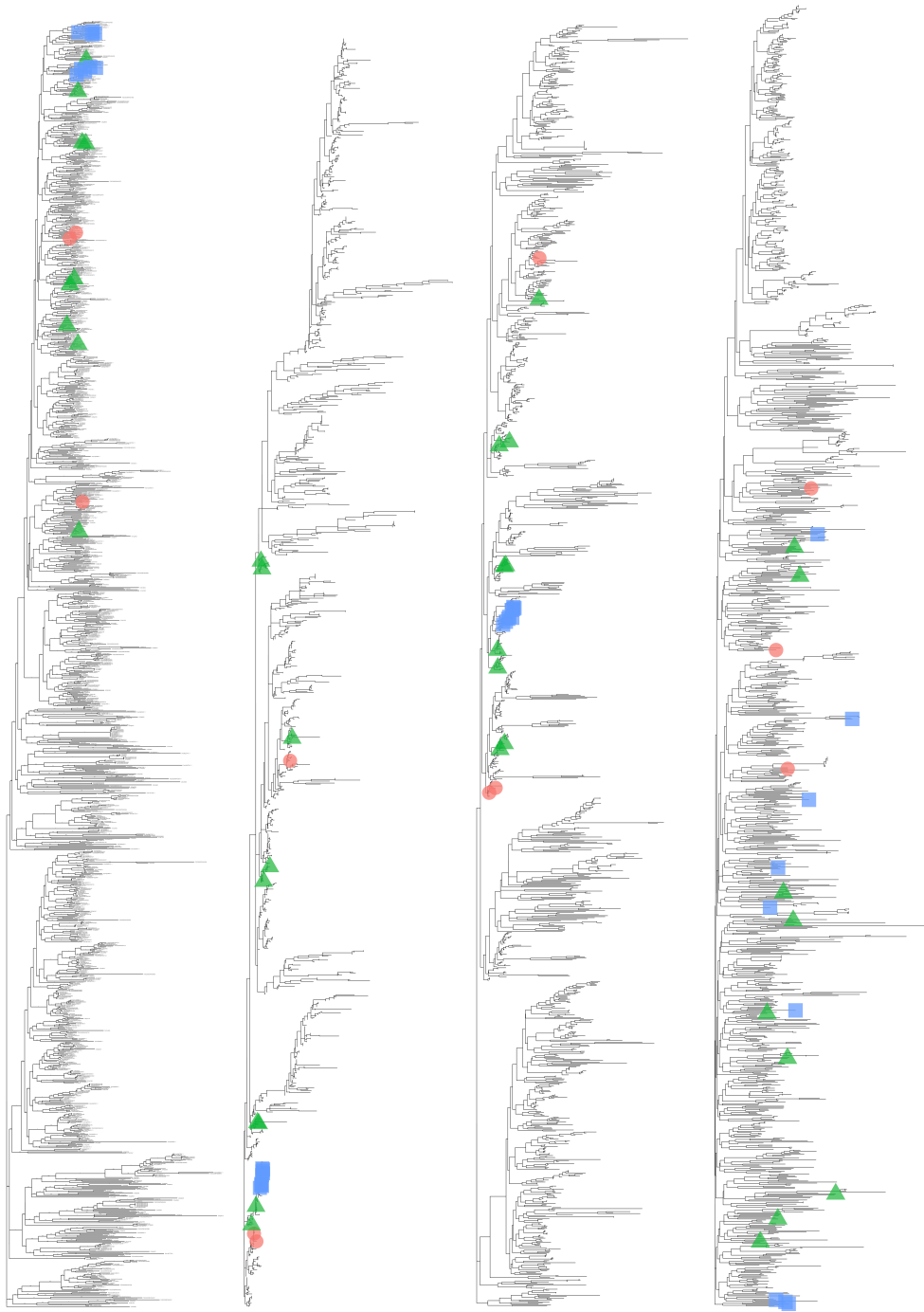
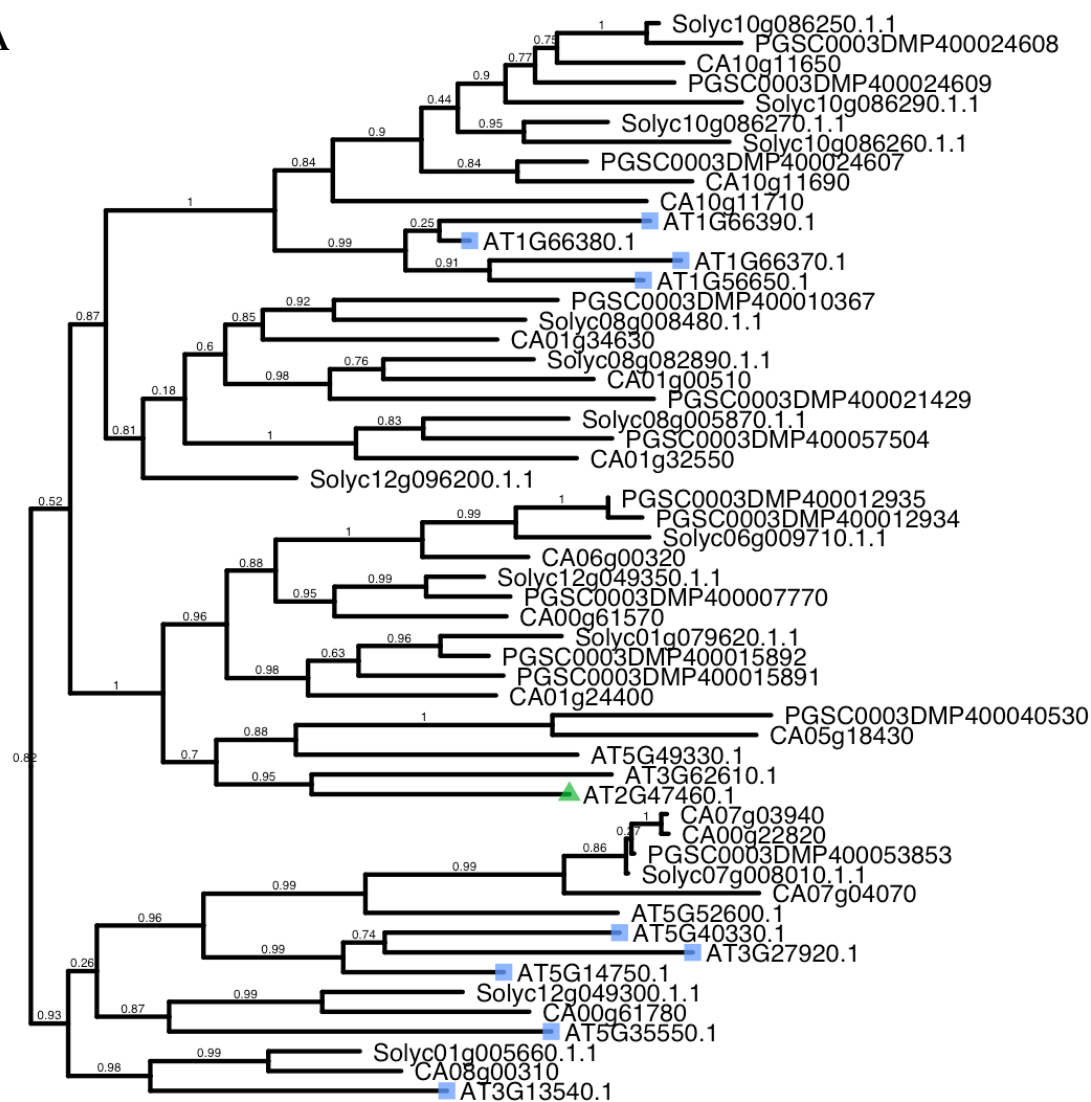


Figure 6. Comparison of Approximately-Maximum Likelihood Phylogenies Generated with Subsets of MYB TF Protein Alignment. A total of 1304 sequences containing MYB motifs from four species were used in the alignment (see Methods in Chapter 7). Protein sequence length was between 52 and 1656 amino acids, with a resulting alignment length of 3492 characters. MYB TFs found to interact with shared protein partners in Zimmermann et al. 2004 are marked with blue squares, MYB TFs found to interact with a shared protein partner in Lumba et al. 2014 are marked with red circles and MYB TFs found to interact with a shared DNA binding partner in Taylor-Teeple et al. 2015 are marked with green triangles. Phylogenies were visualized with ggtree (Yu et al. 2017).

- A. *Left*:** Phylogeny using full MYB TF protein alignment.
- B. *Center Left*:** Phylogeny using an alignment subsection containing the protein-protein interaction motif identified in Zimmerman et al. 2004. This phylogeny was constructed using the region aligned with the 20 aa-long motif described in Zimmermann et al., 2004. The motif [DE]Lx2[RK]x3Lx6Lx3R was located in the alignment, and the alignment subsection was cut 2 amino acids before the start of the motif, and one amino acid following the motif (approximately 23 amino acids selected in TFs that contain the motif). Because of indels in the region, the length of the alignment selected is 200 characters.
- C. *Center Right*:** Phylogeny using the first 2109 characters of the alignment. This section contains all detected MYB motifs.
- D. *Right*:** Phylogeny using characters 2110-3492 of the alignment. This section contained no detected MYB motifs.

A



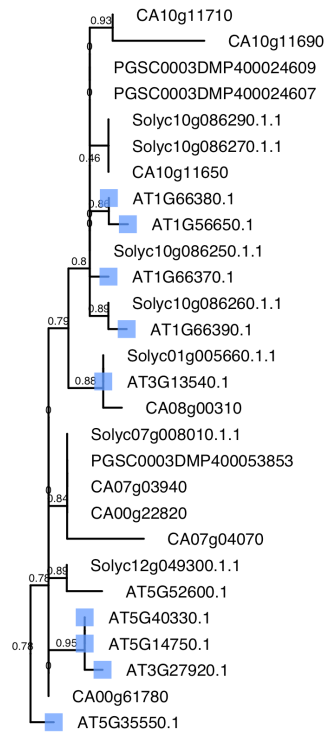
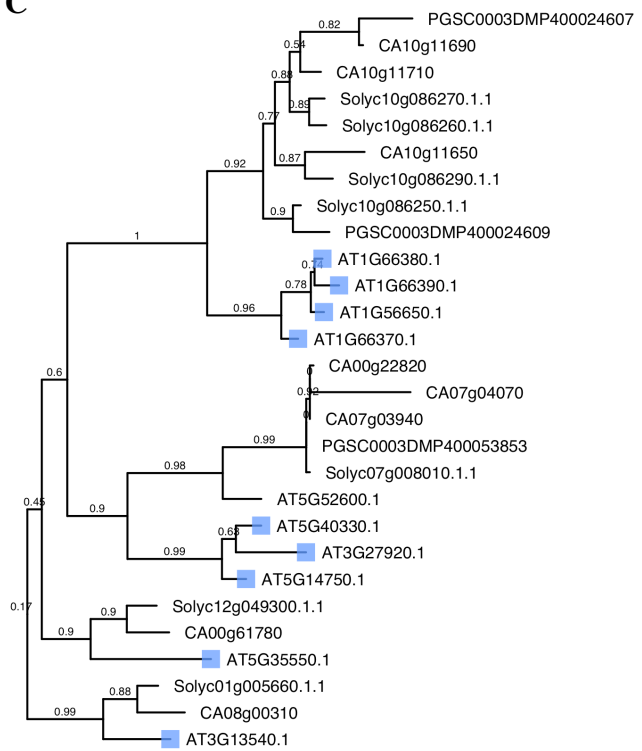
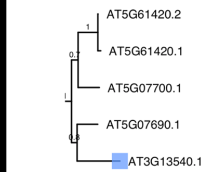
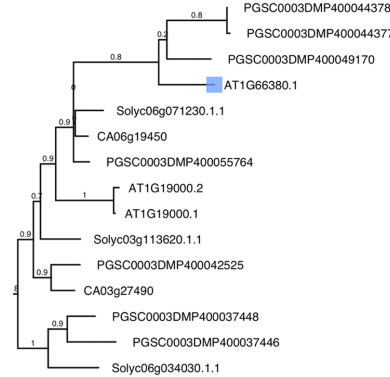
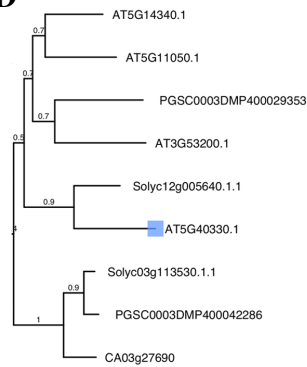
B**C****D**

Figure 7. Detailed Views of Approximately-Maximum Likelihood Phylogenies Generated with Subsets of MYB TF Protein Alignment.

Each phylogeny shown corresponds to the matching letter in Figure 6. Here the clades containing MYB TFs that were characterized in Zimmerman et al. are magnified to provide a view of local support values and tree structure. In the case of the phylogeny using characters 2110-3492 of the alignment (part D), the MYB TFs characterized in Zimmerman et al. 2004 do not corresponded to a single clade. Instead, a subset of the TFs were chosen, and the clade each belonged to is displayed. Phylogenies were visualized with ggtree (Yu et al. 2017).

- A. Phylogeny using full MYB TF protein alignment.**
- B. Phylogeny using alignment subsection containing the protein-protein interaction motif identified in Zimmerman et al. 2004.**
- C. Phylogeny using the first 2109 characters of the MYB TF protein alignment.**
- D. Phylogeny using characters 2110-3492 of the alignment.**

CHAPTER 6

DISCUSSION

The goal of this research was to examine the results of untargeted screens in order to evaluate and improve computational methods for predicting MYB gene function.

Studies were selected that used untargeted methods to identify MYB TFs that interacted with a shared DNA or protein-binding partner. These MYB TFs were mapped on to the widely used classification scheme for MYB genes in Figure 4. Figure 4 shows that the relationships between the MYB TFs that interact with a specific protein or DNA partner vary. MYB TFs that bind the same partner do not cluster together in the phylogeny. Some MYB TFs that interact with the same protein or DNA partner appear to be only distantly related to each other.

This result is expected in cases where the target contains multiple binding sites or when the MYB TFs involved have diverged in activity but compete with one another for binding sites. The pattern of relationships could also indicate recombination or gene conversion within the MYB family. In the case where membership in a shared clade corresponded with similar binding activity (Zimmermann et al. 2004), some genes that shared a common binding partner did not contain any MYB motifs, and so would have been missed by most methodologies used to generate a phylogeny of this type. However, if the authors had searched for genes using sequence similarity in the relevant binding region they likely would have identified the additional factors that

they pulled-down in their untargeted assay. This research suggests that broad searches in a biological context, such as untargeted pull down or hybridization experiments, are necessary to build better models of MYB TF interactions.

In addition to recommending untargeted methods to search for MYB TFs involved in biological processes, this research also provides insight into the mechanisms of MYB gene evolution. Analysis with RDP4 (Figure 5) identified recombination events across the R2 and R3 MYB regions. Recombination resulting in partial gene conversion or in the transfer of sequences to new places in the genome provides an explanation as to why seemingly distant genes could share functions or interaction partners. This result is supported by the phylogenies using subsections of the alignment. The phylogeny of the N-terminal region containing MYB motifs describes a very different set of relationships than the phylogeny of the C-terminal region. Zooming in on specific examples, in panel C of Figure 7 we can see that AT1G66380.1, AT1G66390.1, AT1G56650.1 and AT1G66370.1 form a small clade with good bootstrap support. In the upper right section of panel D AT1G66370.1 and AT1G56650.1 are still together in a strongly supported clade, but the other two genes are now located very distantly in the phylogeny, as we can see on a larger scale in Figure 6. A pattern like this could be explained by recombination of the MYB-containing region.

Recombination also provides a good hypothesis to explain the overall structure of the MYB family. It is clear that MYB genes have proliferated in the genome and that their structures are incredibly diverse. Among the genes pulled-down by Zimmerman et al.

are four genes with either 1 or 0 recognizable MYB motifs that interacted with the same bHLH bait as the genes with full R2R3 MYB motifs. While these genes are very similar in a small portion of their sequence (the protein interaction region), outside of this sequence they bear little similarity to each other. A pattern like this has many possible causes, including gene duplications and later divergence and loss of surrounding sequences. However, in the context of a family with high structural diversity (i.e., the motif may be located almost anywhere along the length of the gene), recombination offers a simple explanation.

Further research is required to validate the recombination hypothesis, as the results shown are not conclusive. Future experiments could combine recombination prediction with data about protein and DNA binding. Consider MYB TFs that contain both a well-characterized protein-binding motif and a well-characterized DNA-binding motif: comparison of the phylogenies of those regions could show that the two binding motifs have been linked throughout their evolution, or that their evolutionary histories are quite different. Current datasets are limited in that most only test protein or DNA interactions, and only across a very small, preselected panel of partners. More data about interactions, especially data that is untargeted and spread evenly across the phylogeny, will be critical to developing predictive techniques.

Without comprehensive data available, it appears that focusing on short specific sequences associated with specific binding activities may be valuable. Improvements in modeling protein folding and binding could contribute to improvements in

prediction of MYB gene function. Alignment-free sequence comparison techniques are becoming increasingly sophisticated and could provide breakthroughs in our ability to understand complex gene families in situations where structural variation undermines traditional alignment techniques (Vinga and Almeida 2003; Leimeister et al. 2014; Thankachan et al. 2017; Cattaneo et al. 2017). The complex evolutionary history of the MYB TF family provides many opportunities for research, with broad implications into the way that many large gene families have evolved.

CHAPTER 7

MATERIALS AND METHODS

Identifying MYB sequences for analysis

MYB genes were identified by BLASTp search against the potato (Potato PGSC DM v3.4 protein sequences), tomato (Tomato Proteins, ITAG release 2.40), pepper (Pepper Genome Protein Sequences, release 1.55) and *Arabidopsis thaliana* (Arabidopsis proteins, TAIR 10) databases, obtained through the solgenomics.net “old” advanced interface on 3/31/2014. The query sequences were obtained from the Plant Transcription Factor database on 3/29/2014. In the Plant Transcription Factor Database, 8746 genes from 83 species were designated as part of the MYB family of Transcription Factors, and all were used as part of the MYB-gene query for BLASTp.

After queries were returned, the resulting matches were filtered to remove duplicates (by accession number) and to create databases for each species. At the end of the process there were 392 MYB genes in potato, 337 MYB genes in tomato, 360 MYB genes in pepper, and 215 MYB genes in *Arabidopsis*.

Nucleotide sequences used for analysis of Arabidopsis MYB genes were found using the protein accession labels to query for the corresponding nucleotide sequence using TAIR (10/8/2015).

The oddZimm set includes 4 genes detected in Zimmermann et al. with their set of bHLH genes as bait. One of these genes, AT171030, contains a single MYB motif. The other genes had no motifs detectable by PROSITE, but do contain sequences similar to MYB motifs. The genes included in this set are AT1G71030 (AtMYBL2), AT2G46410 (CPC), AT1G01380, and AT2G30420.

Alignment

Several alignments were produced over the course of this research (Table 4). Various parameters were tried and assessed by visual inspection of the alignment. Default values were selected if improvements were not visually evident.

Table 4. Alignments used in this research.

Label	Alignment Algorithm	Dataset Aligned	Gap-Opening Penalty	Gap-Extension Penalty
All_MYB_Protein_1	MUSCLE	All MYB protein sequences (above)	-2.9	0
Ara_MYB_Nuc_1	MUSCLE	Arabidopsis MYB nucleotide sequences (above)	-500	-0.5
All_MYB_Protein_oddZimm	MUSCLE	All MYB protein sequences (above) plus the oddZimm set described above	-2.9	0

The software MEGA5.2.2 was used to implement the MUSCLE alignment algorithm.

Phylogeny

RAxML

The phylogeny in Figure 4 was generated using RAxML through the CIPRES portal (Miller, Pfeiffer, and Schwartz 2010; Miller, Pfeiffer, and Schwartz 2011). The tool used was RAxML-HPC2 on XSEDE (version 8.1.24) (Stamatakis 2014). Protein was selected as the sequence type, and the General Time Reversible (GTR) model was selected as the Protein Substitution Matrix. Bootstrap iterations were set to 250 because iterations of 1000 and 500 bootstraps failed to complete within 20,480 cpu hours (in version 8.0.9). Other values were maintained at default settings for that version number.

FastTree

Phylogenies in Figures 6-7 were generated using FastTree (Price, Dehal, and Arkin 2010) in the CIPRES Portal (Miller, Pfeiffer, and Schwartz 2011). Amino Acid was selected as the data type, JTT+CAT was selected as the Substitution Model, and 1000 bootstraps were designated. Other settings were maintained at their default value.

Annotation and Visualization

Annotation and visualization was performed in both FigTree v1.4.2 (Raumbaut, 2014) and ggtree (Yu et al. 2017). Figures were generated in ggtree. Protein interaction data was accessed through direct download of files organized by organism from the BIOGRID 3.4.127 release (compiled 7/25/2015, accessed 8/12/2015). MYB gene

accession numbers used in the phylogeny were used to query the database (Chatr-aryamontri et al. 2015). DNA interaction data was accessed through direct download of files organized by gene ID from the AGRIS AtRegNet database (updated 5/21/2015, accessed 3/16/2016). Genes annotated as ‘MYB’ with confirmed interactions were extracted from the database (Yilmaz et al. 2011).

Recombination Analysis

RDP4 is a group of programs used to detect recombination events. Recombinant regions are identified by comparing gene segments within the alignment. If two genes align with each other very well in just one region, but align best with other genes when considering sequences outside that region, this may be because of recombination. Certainty (given as p-values, and calculated differently depending on the detection method) is measured by estimating the likelihood of detecting these differences by chance.

Arabidopsis nucleotide alignments (Ara_MYB_Nuc1, Table 3) were scanned for recombination events using Recombination Detection Program 4 (RDP4). Nucleotide sequences from *Arabidopsis* were used because RDP4 cannot process protein sequences. For Figure 5 default settings for RDP4 were used with the alignment Ara_MYB_Nuc1, with selections made to indicate linear DNA sequences. This included the analysis tools RDP, GENECONV, MaxChi, Bootscan and SiScan.

REFERENCES

- Bedon, Frank, Claude Bomal, Sebastien Caron, Caroline Levasseur, Brian Boyle, Shawn D. Mansfield and Axel Schmidt. 2010. "Subgroup 4 R2R3-MYBs in Conifer Trees: Gene Family Expansion and Contribution to the Isoprenoid-and Flavonoid-Oriented Responses." *Journal of Experimental Botany* 61 (14): 3847–64. doi:10.1093/jxb/erq196.
- Blanc, G, A Barakat, R Guyot, R Cooke, and M Delseny. 2000. "Extensive Duplication and Reshuffling in the Arabidopsis Genome." *The Plant Cell* 12 (7): 1093–1101. doi:10.1105/tpc.12.7.1093.
- Cannon, Steven B, Arvind Mitra, Andrew Baumgarten, Nevin D Young, and Georgiana May. 2004. "The Roles of Segmental and Tandem Gene Duplication in the Evolution of Large Gene Families in Arabidopsis Thaliana." *BMC Plant Biology* 4: 10. doi:10.1186/1471-2229-4-10.
- Cao, Zhong-Hui, Shi-Zhong Zhang, Rong-Kai Wang, Rui-Fen Zhang, and Yu-Jin Hao. 2013. "Genome Wide Analysis of the Apple MYB Transcription Factor Family Allows the Identification of MdoMYB121 Gene Confering Abiotic Stress Tolerance in Plants." *PloS One* 8 (7): e69955. doi:10.1371/journal.pone.0069955.
- Cattaneo, Giuseppe, Umberto Ferraro Petrillo, Raffaele Giancarlo, and Gianluca Roscigno. 2017. "An Effective Extension of the Applicability of Alignment-free Biological Sequence Comparison Algorithms with Hadoop." *J Supercomput* 73: 1467-1483. doi: 10.1007/s11227-016-1835-3.

- Chatr-aryamontri, A, BJ Breitkreutz, R Oughtred, L Boucher, S Heinicke, D Chen and C Stark. 2015. “The BioGRID Interaction Database: 2015 Update.” *Nucleic Acids Research* 43 (D1): D470–78. doi:10.1093/nar/gku1204.
- Du, Hai, Bo-Run Feng, Si-Si Yang, Yu-Bi Huang and Yi-Xiong Tang. 2012. “The R2R3-MYB Transcription Factor Gene Family in Maize.” Edited by Keqiang Wu. *PLoS ONE* 7 (6): e37463. doi:10.1371/journal.pone.0037463.
- Du, Hai, Yongin Wang, Yi Xie, Zhe Liang, Sanie Jiang, S Huang Huang Zhang and Yui Huang. 2013. “Genome-Wide Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants” 1 (May): 437–48.
- Du, Hai, Si-Si Yang, Zhe Liang, Bo-Run Feng, Lei Liu, Yu-Bi Huang, and Yi-Xiong Tang. 2012. “Genome-Wide Analysis of the MYB Transcription Factor Superfamily in Soybean.” *BMC Plant Biology* 12 (1): 106. doi:10.1186/1471-2229-12-106.
- Dubos, Christian, Ralf Stracke, Erich Grotewold, Bernd Weisshaar, Cathie Martin, and Loïc Lepiniec. 2010. “MYB Transcription Factors in *Arabidopsis*.” *Trends in Plant Science* 15 (10): 573–81. doi:10.1016/j.tplants.2010.06.005.
- Feldbrügge, Michael, Markus Sprenger, Klaus Hahlbrock, and Bernd Weisshaar. 1997. “PcMYB1, a Novel Plant Protein Containing a DNA-Binding Domain with One MYB Repeat, Interacts in Vivo with a Light-Regulatory Promoter Unit.” *Plant Journal* 11 (5): 1079–93. doi:10.1046/j.1365-313X.1997.11051079.x.
- Feller, Antje, Katja MacHemer, Edward L. Braun, and Erich Grotewold. 2011.

- “Evolutionary and Comparative Analysis of MYB and bHLH Plant Transcription Factors.” *Plant Journal* 66 (1): 94–116. doi:10.1111/j.1365-313X.2010.04459.x.
- Feller, Antje, Katja Machemer, Edward L Braun, and Erich Grotewold. 2011.
- “Evolutionary and Comparative Analysis of MYB and bHLH Plant Transcription Factors.” *The Plant Journal : For Cell and Molecular Biology* 66 (1): 94–116. doi:10.1111/j.1365-313X.2010.04459.x.
- Frerigmann, Henning, Bettina Berger, and Tamara Gigolashvili. 2014. “bHLH05 Is an Interaction Partner of MYB51 and a Novel Regulator of Glucosinolate Biosynthesis in Arabidopsis.” *Plant Physiology* 166 (1): 349–69. doi:10.1104/pp.114.240887.
- Hou, Xiao-Jin, Si-Bei Li, Sheng-Rui Liu, Chun-Gen Hu, and Jin-Zhi Zhang. 2014.
- “Genome-Wide Classification and Evolutionary and Expression Analyses of Citrus MYB Transcription Factor Families in Sweet Orange.” *PLoS ONE* 9 (11): e112375. doi:10.1371/journal.pone.0112375.
- Huson, D H. 1998. “SplitsTree: Analyzing and Visualizing Evolutionary Data.” *Bioinformatics (Oxford, England)* 14 (1): 68–73. doi:btb043 [pii].
- Hwang, Moo Gak, In Kwon Chung, Bin Goo Kang, and Myeon Haeng Cho. 2001.
- “Sequence-Specific Binding Property of Arabidopsis Thaliana Telomeric DNA Binding Protein 1 (AtTBP1).” *FEBS Letters* 503 (1): 35–40. doi:10.1016/S0014-5793(01)02685-0.
- Jia, Li, Michael T Clegg, and Tao Jiang. 2004. “Evolutionary Dynamics of the DNA-

Binding Domains in Putative R2R3-MYB Genes Identified from Rice Subspecies Indica and Japonica Genomes.” *Plant Physiology* 134 (2): 575–85.

doi:10.1104/pp.103.027201.

Katiyar, Amit, Shuchi Smita, Sangram Keshari Lenka, Ravi Rajwanshi, Viswanathan Chinnusamy, and Kailash Chander Bansal. 2012. “Genome-Wide Classification and Expression Analysis of MYB Transcription Factor Families in Rice and Arabidopsis.” *BMC Genomics* 13 (January): 544. doi:10.1186/1471-2164-13-544.

Klempnauer, Karl Heinz, Thomas J. Gonda, and J. Michael Bishop. 1982. “Nucleotide Sequence of the Retroviral Leukemia Gene v-Myb and Its Cellular Progenitor c-Myb: The Architecture of a Transduced Oncogene.” *Cell* 31 (2 PART 1): 453–63. doi:10.1016/0092-8674(82)90138-6.

Koshino-Kimura, Yoshihiro, Takuji Wada, Tatsuhiko Tachibana, Ryuji Tsugeki, Sumie Ishiguro, and Kiyotaka Okada. 2005. “Regulation of CAPRICE Transcription by MYB Proteins for Root Epidermis Differentiation in Arabidopsis.” *Plant and Cell Physiology* 46 (6): 817–26. doi:10.1093/pcp/pci096.

Kranz, H D, M Denekamp, R Greco, H Jin, a Leyva, R C Meissner, K Petroni, et al. 1998. “Towards Functional Characterisation of the Members of the R2R3-MYB Gene Family from Arabidopsis Thaliana.” *The Plant Journal : For Cell and Molecular Biology* 16 (2): 263–76.

<http://www.ncbi.nlm.nih.gov/pubmed/9839469>.

- Leimeister, Chris-Andre, Marcus Boden, Sebastian Horwege, Sebastian Lindner, Burkhard Morgenstern. 2014. “Fast alignment-free sequence comparison using spaced-word frequencies.” *Bioinformatics* 30 (14): 1991–1999. doi:10.1093/bioinformatics/btu177.
- Liao, Yong, Hong-Feng Zou, Hui-Wen Wang, Wan-Ke Zhang, Biao Ma, Jin-Song Zhang, and Shou-Yi Chen. 2008. “Soybean GmMYB76, GmMYB92, and GmMYB177 Genes Confer Stress Tolerance in Transgenic Arabidopsis Plants.” *Cell Research* 18: 1047–60. doi:10.1038/cr.2008.280.
- Lu, Chung-An, Tuan-hua David Ho, Shin-Lon Ho, and Su-May Yu. 2002. “Three Novel MYB Proteins with One DNA Binding Repeat Mediate Sugar and Hormone Regulation of Alpha-Amylase Gene Expression.” *The Plant Cell* 14 (8): 1963–80. doi:10.1105/tpc.001735.
- Lumba, Shelley, Shigeo Toh, Louis-François Handfield, Michael Swan, Raymond Liu, Ji-Young Youn, Sean R. Cutler, et al. 2014. “A Mesoscale Absciscic Acid Hormone Interactome Reveals a Dynamic Signaling Landscape in Arabidopsis.” *Developmental Cell* 29 (3): 360–72. doi:10.1016/j.devcel.2014.04.004.
- Martin, D P, B Murrell, M Golden, A Khoosal, and B Muhire. 2015. “RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes.” *Virus Evolution* 1 (1): 1–5. doi:10.1093/ve/vev003.
- Matus, José Tomás, Felipe Aquea, and Patricio Arce-Johnson. 2008. “Analysis of the Grape MYB R2R3 Subfamily Reveals Expanded Wine Quality-Related Clades and Conserved Gene Structure Organization across Vitis and Arabidopsis

- Genomes.” *BMC Plant Biology* 8 (January): 83. doi:10.1186/1471-2229-8-83.
- Miller, Mark A., Wayne Pfeiffer, and Terri Schwartz. 2010. “Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees.” *2010 Gateway Computing Environments Workshop (GCE)*, November. Ieee, 1–8. doi:10.1109/GCE.2010.5676129.
- Miller, Mark A, Wayne Pfeiffer, and Terri Schwartz. 2011. “The CIPRES Science Gateway : A Community Resource for Phylogenetic Analyses.” *2011 TeraGrid Conference: Extreme Digital Discovery*. doi:10.1145/2016741.2016785.
- Montefiori, M., C. Brendolise, a. P. Dare, K. Lin-Wang, K. M. Davies, R. P. Hellens, and a. C. Allan. 2015. “In the Solanaceae, a Hierarchy of bHLHs Confer Distinct Target Specificity to the Anthocyanin Regulatory Complex.” *Journal of Experimental Botany*. doi:10.1093/jxb/eru494.
- Ogata, Kazuhiro, Chie Kanei-Ishii, Motoko Sasaki, Hideki Hatanaka, Aritaka Nagadoi, Masato Enari, Haruki Nakamura, Yoshifumi Nishimura, Shunsuke Ishii, and Akinori Sarai. 1996. “The Cavity in the Hydrophobic Core of Myb DNA-Binding Domain Is Reserved for DNA Recognition and Trans-Activation.” *Nat Struct Mol Biol* 3 (2): 178–87. <http://dx.doi.org/10.1038/nsb0296-178>.
- Ogata, Kazuhiro, Souichi Morikawa, Haruki Nakamura, Ai Sekikawa, Taiko Inoue, Hiroko Kanai, Akinori Sarai, Shunsuke Ishii, and Yoshifumi Nishimura. 1994. “Solution Structure of a Specific DNA Complex of the Myb DNA-Binding Domain with Cooperative Recognition Helices.” *Cell* 79 (4): 639–48. doi:10.1016/0092-8674(94)90549-5.

- Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS ONE* 5 (3). doi:10.1371/journal.pone.0009490.
- Prouse, Michael B., and Malcolm M. Campbell. 2012. "The Interaction between MYB Proteins and Their Target DNA Binding Sites." *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1819 (1). Elsevier B.V.: 67–77. doi:10.1016/j.bbagr.2011.10.010.
- Rabinowicz, Pablo D., Edward L. Braun, Andrea D. Wolfe, Ben Bowen, and Erich Grotewold. 1999. "Maize R2R3 Myb Genes: Sequence Analysis Reveals Amplification in the Higher Plants." *Genetics* 153 (1): 427–44.
- Rambaut, Andrew. 2014. "FigTree v1.4.2." <http://tree.bio.ed.ac.uk/software/figtree/>.
- Riechmann, J L, J Heard, G Martin, L Reuber, C Jiang, J Keddle, L Adam, et al. 2000. "Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis among Eukaryotes." *Science (New York, N.Y.)* 290 (5499): 2105–10. doi:10.1126/science.290.5499.2105.
- Romero, I., A. Fuertes, M. J. Benito, J. M. Malpica, A. Leyva, and J. Paz-Ares. 1998. "More than 80R2R3-MYB Regulatory Genes in the Genome of Arabidopsis Thaliana." *Plant Journal* 14 (3): 273–84. doi:10.1046/j.1365-313X.1998.00113.x.
- Seoighe, C, and C Gehring. 2004. "Genome Duplication Led to Highlyselective Expansion of the Arabidopsis Thaliana Proteome." *Trends Genet* 20 (10): 461–

64.

Serpa, Viviane, Javier Vernal, Lorenzo Lamattina, Erich Grotewold, Raul Cassia, and Hernán Terenzi. 2007. "Inhibition of AtMYB2 DNA-Binding by Nitric Oxide Involves Cysteine S-Nitrosylation." *Biochemical and Biophysical Research Communications* 361 (4): 1048–53. doi:10.1016/j.bbrc.2007.07.133.

Shapiro, Harris, Tomoaki Nishiyama, Pierre-françois Perroud, and Erika a Lindquist. 2008. "Conquest of Land by Plants." *Science* 319 (January): 64–69. doi:10.1126/science.1150646.

Shiu, Shin-Han, Ming-Che Shih, and Wen-Hsiung Li. 2005. "Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals." *Plant Physiology* 139 (1): 18–26. doi:10.1104/pp.105.065110.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. doi:10.1093/bioinformatics/btu033.

Stracke, Ralf, Daniela Holtgräwe, Jessica Schneider, Boas Pucker, Thomas Rosleff Sörensen, and Bernd Weisshaar. 2014. "Genome-Wide Identification and Characterisation of R2R3-MYB Genes in Sugar Beet (*Beta Vulgaris*)." *BMC Plant Biology* 14 (1): 249. doi:10.1186/s12870-014-0249-8.

Stracke, Ralf, Martin Werber, and Bernd Weisshaar. 2001. "The R2R3 - MYB Gene Family in *Arabidopsis Thaliana*." *Current Opinion in Plant Biology*, 447–56.

Taylor-Teeple, M, L Lin, M De Lucas, G Turco, T W Toal, A Gaudinier, N F Young,

- et al. 2015. “An Arabidopsis Gene Regulatory Network for Secondary Cell Wall Synthesis.” *Nature* 517 (7536). Nature Publishing Group: 571–75.
doi:10.1038/nature14099.
- Thankachan, Sharma V, Sriram P. Chockalingam, Yongchao Liu, Ambujam and Srinivas Aluru. 2017. “A greedy alignment-free distance estimator for phylogenetic inference.” *BMC Bioinformatics* 18(Suppl 8): 238.
doi 10.1186/s12859-017-1658-0.
- Vinga, Susana, and Jonas Almeida. 2003. “Alignment-Free Sequence Comparison - A Review.” *Bioinformatics* 19 (4): 513–23. doi:10.1093/bioinformatics/btg005.
- Woodhouse, Margaret R., Brent Pedersen, and Michael Freeling. 2010. “Transposed Genes in Arabidopsis Are Often Associated with Flanking Repeats.” *PLoS Genetics* 6 (5): e1000949. doi:10.1371/journal.pgen.1000949.
- Yanhui, Chen, Yang Xiaoyuan, He Kun, Liu Meihua, Li Jigang, Gao Zhaofeng, Lin Zhiqiang, et al. 2006. “The MYB Transcription Factor Superfamily of Arabidopsis: Expression Analysis and Phylogenetic Comparison with the Rice MYB Family.” *Plant Molecular Biology* 60 (1): 107–24. doi:10.1007/s11103-005-2910-y.
- Yilmaz, Alper, Maria Katherine Mejia-Guerra, Kyle Kurz, Xiaoyu Liang, Lonnie Welch, and Erich Grotewold. 2011. “AGRIS: The Arabidopsis Gene Regulatory Information Server, an Update.” *Nucleic Acids Research* 39 (SUPPL. 1): 1118–22. doi:10.1093/nar/gkq1120.

- Yu, Guangchuang, David K Smith, Huachen Zhu, Yi Guan and Tommy Tsan-Yuk Lam. 2017. “ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.” *Methods in Ecology and Evolution* 8: 28–36. doi:10.1111/2041-210X.12628
- Zhao, Lei, Liping Gao, Hongxue Wang, Xiaotian Chen, Yunsheng Wang, Hua Yang, Chaoling Wei, Xiaochun Wan, and Tao Xia. 2013. “The R2R3-MYB, bHLH, WD40, and Related Transcription Factors in Flavonoid Biosynthesis.” *Functional & Integrative Genomics* 13 (1): 75–98. doi:10.1007/s10142-012-0301-4.
- Zhou, Changpin, Yanbo Chen, Zhenying Wu, Wenjia Lu, Jinli Han, Pingzhi Wu, Yaping Chen, Meiru Li, Huawu Jiang, and Guojiang Wu. 2015. “Genome-Wide Analysis of the MYB Gene Family in Physic Nut (*Jatropha Curcas* L.).” *Gene* 572 (1). Elsevier B.V.: 63–71. doi:10.1016/j.gene.2015.06.072.
- Zimmermann, Ilona M, Marc A Heim, Bernd Weisshaar, and Joachim F Uhrig. 2004. “Comprehensive Identification of *Arabidopsis Thaliana* MYB Transcription Factors Interacting with R/B-like BHLH Proteins.” *The Plant Journal : For Cell and Molecular Biology* 40 (1): 22–34. doi:10.1111/j.1365-3113X.2004.02183.x.